Varianza y desviación estándar muestral en Python

Importamos las librerías necesarias para análisis

```
import os

def import_or_install(package):
    try:
        __import__(package)
    except ImportError:
        os.system("pip install "+ package)
```

```
import_or_install('pandas')
import_or_install('numpy')
import_or_install('matplotlib')
import_or_install('seaborn')
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

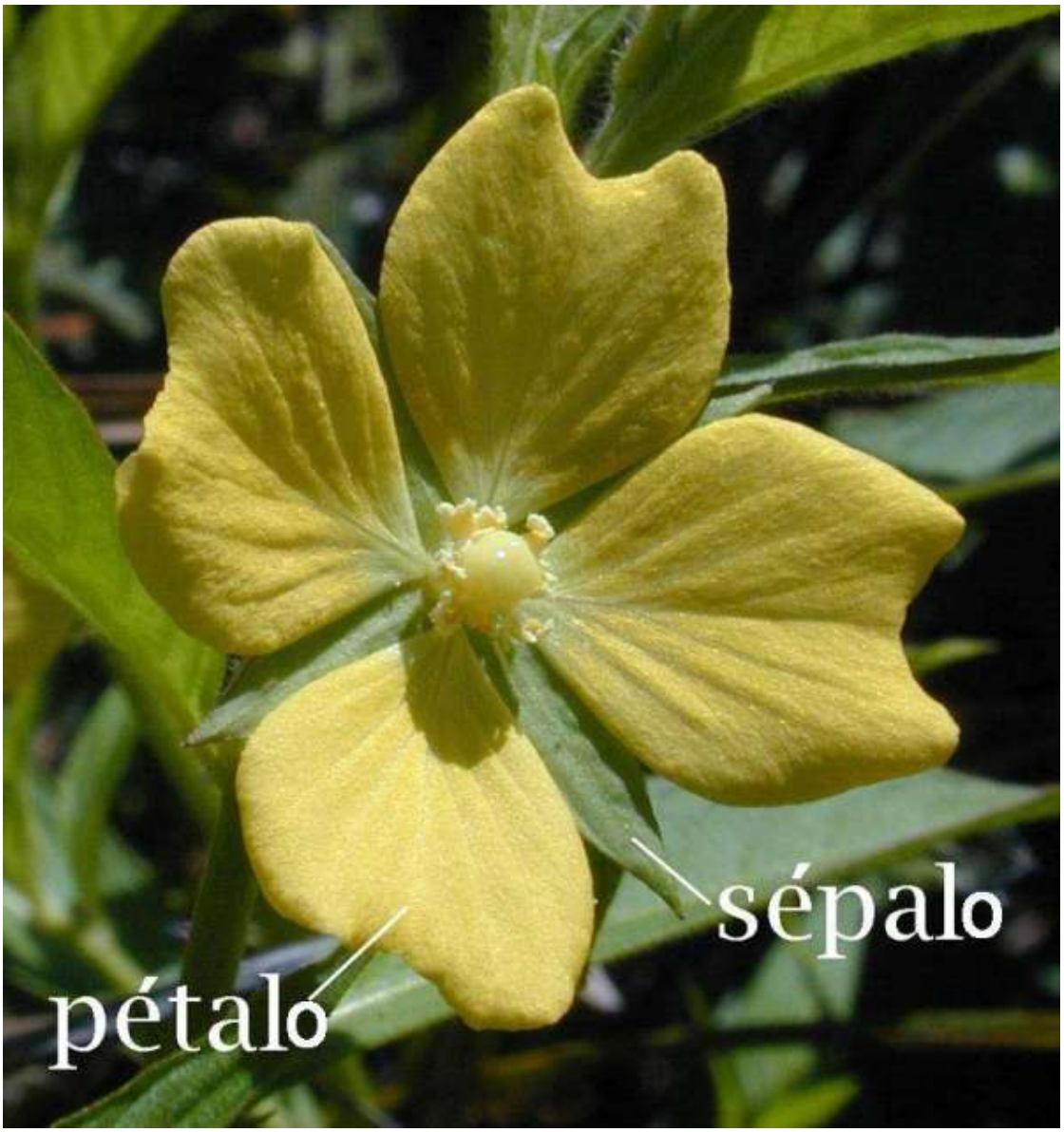
Leemos la base de datos a trabajar en nuestra ruta de origen

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
iris = pd.read_csv(url, names = names)
```

```
iris.head()
```

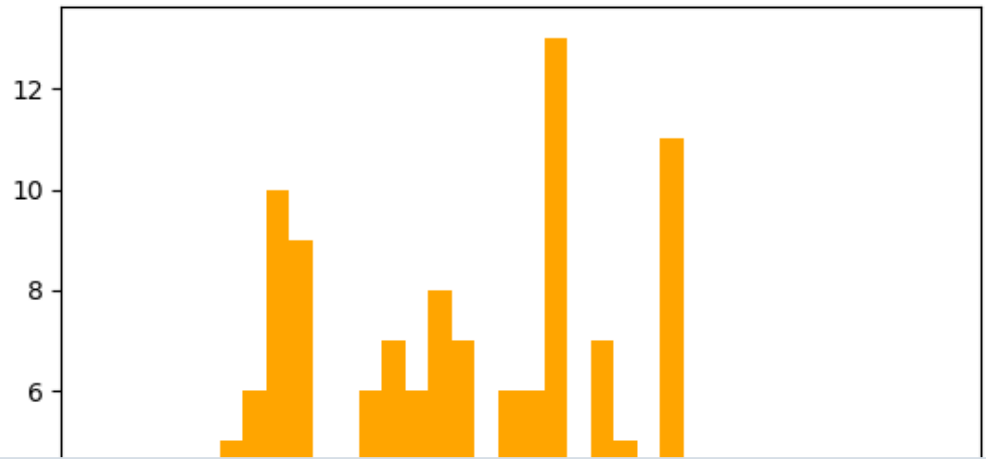
	sepal-length float64	sepal-width float64	petal-length float64	petal-width float64	class object
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5	3.6	1.4	0.2	Iris-setosa

5 rows, showing 10 per page Page 1 of 1



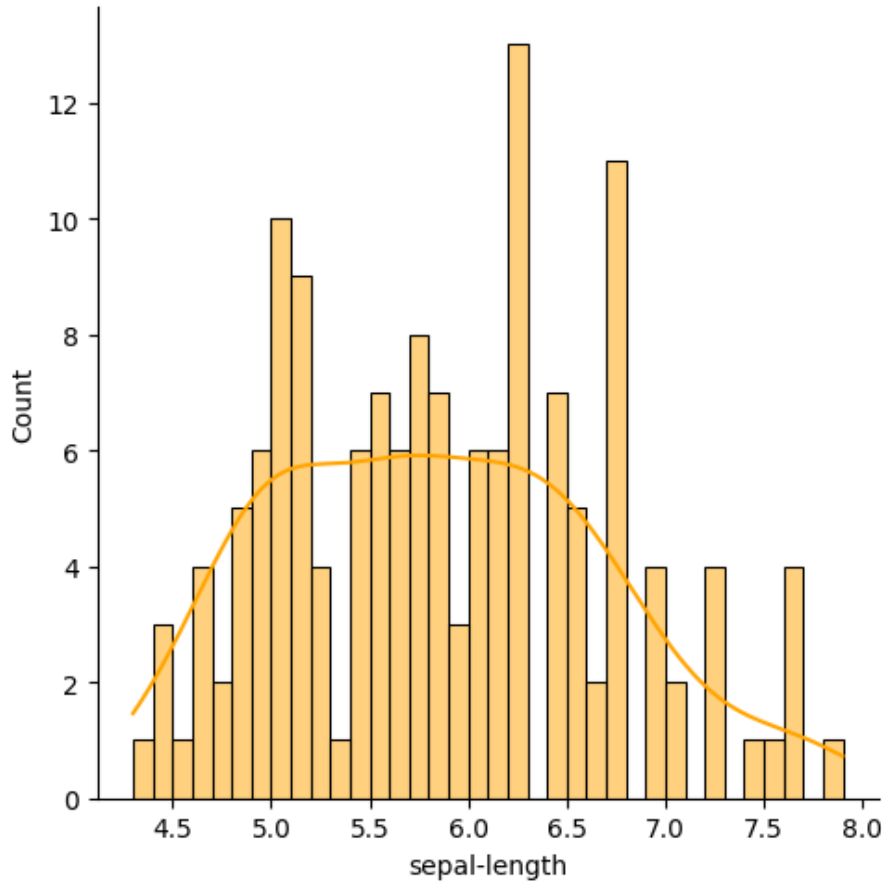

```
plt.hist(iris['sepal-length'], color='orange', bins = int(180/5))
```

```
(array([ 1.,  3.,  1.,  4.,  2.,  5.,  6., 10.,  9.,  4.,  1.,  6.,  7.,
         6.,  8.,  7.,  3.,  6.,  6., 13.,  0.,  7.,  5.,  2., 11.,  0.,
         4.,  2.,  0.,  4.,  0.,  1.,  1.,  4.,  0.,  1.]),
 array([4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5. , 5.1, 5.2, 5.3, 5.4, 5.5,
        5.6, 5.7, 5.8, 5.9, 6. , 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8,
        6.9, 7. , 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7, 7.8, 7.9]),
 <BarContainer object of 36 artists>)
```




```
sns.displot(iris['sepal-length'], kde=True, bins=int(180/5),
            color='orange')
```

```
<seaborn.axisgrid.FacetGrid at 0x7f19af57ffa0>
```



Cálculo de la varianza

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

```
iris['sepal-length'].var()
```

```
0.6856935123042507
```

Cálculo de la desviación estandar

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

```
iris['sepal-length'].std()
```

```
0.828066127977863
```

Cálculo del promedio

```
iris['sepal-length'].mean()
```

```
5.843333333333334
```

La desviación es elevada

Interpretación del coeficiente de variación

CV	Apreciación de la muestra
0 % a 10 %	Muy homogénea
11 % a 15 %	Homogénea
16 % a 25 %	Heterogénea
26 % a más	Muy heterogénea

Creación de una muestra mediante el muestreo aleatorio simple

```
muestra = iris.sample(frac=0.5)
muestra.head()
```

	sepal-length float64	sepal-width float64	petal-length float64	petal-width float64	class object	
13	4.3	3	1.1	0.1	Iris-setosa	
124	6.7	3.3	5.7	2.1	Iris-virginica	
45	4.8	3	1.4	0.3	Iris-setosa	
125	7.2	3.2	6	1.8	Iris-virginica	
10	5.4	3.7	1.5	0.2	Iris-setosa	

5 rows, showing 10 per page Page 1 of 1

Cálculo de la varianza

```
muestra['sepal-length'].var()
```

```
0.7330630630630632
```

Cálculo de la desviación estandar

```
muestra['sepal-length'].std()
```

```
0.8561910201952969
```

Cálculo del promedio

```
muestra['sepal-length'].mean()
```

```
5.926666666666666
```

Conclusiones

Un valor cercano a 1 significa una varianza y desviación elevadas

Desv.Est.

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos alrededor de la media. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido.

Debido a que la desviación estándar utiliza las mismas unidades que los datos, generalmente es más fácil de interpretar que la varianza.

Interpretación

Utilice la desviación estándar para determinar qué tan dispersos están los datos con respecto a la media. Un valor de desviación estándar más alto indica una mayor dispersión de los datos. Una buena regla empírica para una distribución normal es que aproximadamente 68% de los valores se ubican dentro de una desviación estándar de la media, 95% de los valores se ubican dentro de dos desviaciones estándar y 99.7% de los valores se ubican dentro de tres desviaciones estándar.

La desviación estándar también se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.

Varianza

La varianza mide qué tan dispersos están los datos alrededor de su media. La varianza es igual a la desviación estándar elevada al cuadrado.

Interpretación

Mientras mayor sea la varianza, mayor será la dispersión de los datos.

Puesto que la varianza (σ2) es una cantidad elevada al cuadrado, sus unidades también están elevadas al cuadrado, lo que puede dificultar el uso de la varianza en la práctica. La desviación estándar generalmente es más fácil de interpretar porque utiliza las mismas unidades que los datos. Por ejemplo, una muestra del tiempo de espera en una parada de

autobuses puede tener una media de 15 minutos y una varianza de 9 minutos2. Debido a que la varianza no está en las mismas unidades que los datos, la varianza suele mostrarse con su raíz cuadrada, la desviación estándar. Una varianza de 9 minutos2 es equivalente a una desviación estándar de 3 minutos.

```
iris.describe()
```

	sepal-length float64	sepal-width float64	petal-length float64	petal-width float64
cou…	150	150	150	150
me…	5.843333333	3.054	3.758666667	1.198666667
std	0.828066128	0.4335943114	1.76442042	0.7631607417
min	4.3	2	1	0.1
25%	5.1	2.8	1.6	0.3
50%	5.8	3	4.35	1.3
75%	6.4	3.3	5.1	1.8
max	7.9	4.4	6.9	2.5

8 rows, showing 10 per page Page 1 of 1